

White Paper Report

Report ID: 111862

Application Number: HD-51828-14

Project Director: Patrick Manning (pmanning@pitt.edu)

Institution: University of Pittsburgh

Reporting Period: 7/1/2014-6/30/2015

Report Due: 9/30/2015

Date Submitted: 11/10/2015

“World-Historical Gazetteer”

Report type: White Paper

Grant Number: HD-51828-14

Title of Project: World Historical Gazetteer

Name of Project Directors: Patrick Manning (Pitt) and Ruth Mostern (UC-Merced), Co-PIs

Name of grantee institution: University of Pittsburgh

Abstract:

The project for a World-Historical Gazetteer is a cross-disciplinary collaboration of researchers and developers from various academic and professional institutions that are working towards creating a comprehensive, online gazetteer of world-historical places for the period after 1500 CE. This paper describes the main objectives of the project, outlines several different work plans, and reports on developments from the World-Historical Gazetteer Workshop that took place in September of 2014. During the workshop, participants worked to establish a criteria and standards for a world-historical gazetteer and reported on currently available gazetteer resources. A consensus arose among participants to begin the work of developing a “spine” gazetteer, based on a world-historical atlas, which is to be linked to other gazetteer resources. While the initial development work on the spine has already commenced, additional activities will continue through December of 2015. The product of the start-up project is displayed on the CHIA website: a listing of core resources, the initial spine gazetteer itself, and the plan for constructing the final product, the World-Historical Gazetteer. The plan includes steps for linking the spine to an expanding range of library and historical resources distinguishing the general gazetteer from the gazetteer of administrative units. The final product, which will require additional implementation work to complete, will be an open source, linked data resource that is supported by the technical infrastructure of the Collaborative for Historical Information and Analysis. It is expected to become a centerpiece in spatial ontology that can facilitate consistent spatial documentation of historical datasets and lay the groundwork for parallel work in temporal, topical, and scalar ontologies for documenting historical datasets.

“World-Historical Gazetteer”

Patrick Manning (Pitt) and Ruth Mostern (UC-Merced), Co-PIs

White Paper

The project and its objectives

This DH Start-up project on the World-Historical Gazetteer continued for one year. Its objectives were formulated and have been evaluated in three time frames. First was the time frame of a two-day workshop that took place at the beginning of the project year in September 2014. Second was the one-year time frame for pursuing the insights arising from the workshop. Third is the set of hopes and plans for longer-term implementation and effects of the initial plan.

The project’s final objective, as given in the proposal, is a comprehensive, online gazetteer of world-historical places for the period after 1500 CE. More specific objectives are (A) a set of standards and criteria for worldwide documentation of places, (B) a database including a large quantity of consistently formatted specific data on places, drawn from research of participants and from other resources, (C) a system of ingest for continuing incorporation of additional gazetteer data, and (D) a user interface enabling researchers to apply the gazetteer. The plan of work emphasized (1) linking place names from Wikidata, Library of Congress, and other resources; (2) building an inventory of existing gazetteers; (3) forming a strategy for a high-level gazetteer; (4) identifying potential funding sources for continuing work.

The project for a World-Historical Gazetteer (WHG)—based at the World History Center of the University of Pittsburgh and closely linked to the Collaborative for Historical Information and Analysis (CHIA)—began its work in September 2014. The WHG participants, roughly twenty in number, are drawn from the Digital Humanities Geohumanities group, the CHIA project, Pelagios (Pleiades and China Historical GIS), Past Place, and the New York Public Library. In disciplinary terms, they are developers, spatial analysts, librarians, historians, and social scientists, with expertise in digital humanities, big data, and gazetteers.

Opening Workshop and its Activities

Project activity began with a two-day workshop of the twenty participants, Sept. 4-5, 2014, emphasizing the strategy for building a high-level gazetteer. In response to an agenda proposed by the workshop convenors, the group adopted the notion of a simplified database—a “spine”—listing places relevant in world history over the past five centuries, to which other and more specific gazetteers are to be linked. The gazetteer is to be flexible, including places of recent times that are fully documented, but also places from earlier

times that are known to have been important but may not be fully documented.

The workshop began with participant reports on their current work including updates from the Digital Humanities Conference in Lausanne and an introduction to the Pelagios Project that enables linked ancient geodata in open systems. In addition, participants learned about cross-gazetteer search frameworks and gazetteer-interlinking RDF profiles as reported by the Pelagios Project. Participants were also introduced to the PastPlace project as a possible prototype world-historical gazetteer. There followed an open-ended discussion on specifications of a world-historical gazetteer, within four groups, which reported to the overall workshop. This resulted in a common statement of criteria for a world-historical gazetteer, summarized in the next section. The group then worked as a committee of the whole in a design session for the gazetteer, led by Raj Singh. This design session resulted in agreement on a "spine" of a gazetteer, to consist of world-historical places. That is, the initial gazetteer is to be a relatively simple list of key historical places—a "spine"—to which other spatial data and other resources are to be attached. While the spine may be described as a "simple" list, the discussion generated the potential for a very complex set of connections that would ultimately be included in the gazetteer to provide breadth of coverage and depth of details. The workshop continued with discussion and basic agreement on the types of resources to be attached to the spine, a plan for creation and implementation of the spine, and discussion of the additional steps for further materials to be linked to the spine.

Criteria and Standards for the World-Historical Gazetteer. Since no existing gazetteer can serve as a basis for a post-1500 world-historical gazetteer, it is necessary to build a new one. The new gazetteer is to adopt, in large part, the strategy, tactics, and tools developed for pre-1500 gazetteers by Pelagios and Pleiades projects. The world-historical gazetteer is to be a "spine" gazetteer to which other resources will be linked. The gazetteer:

1. Is to be based on open data and linked open data;
2. Is to be simple and basic — others will link elements to it to expand and create more complex, specialized gazetteers;
3. Its content is to include a wide range of places, including places for which there is no information on name, location, or feature type, but for which there is a desire to fill in the missing information;
4. Is to allow specification of relationships among places, both hierarchical and parallel;
5. Is to be linked to a worldwide gazetteer of administrative units;
6. Is to be aligned with the world-historical goals of CHIA, and it is to give special attention to coverage of Africa, Latin America, and Southeast Asia.

At the conclusion of the workshop, participants proposed relevant follow-up activities to be completed within the following year, specifying which individuals would work on each. The spine gazetteer was to be created from relevant resources. Initial discussion centered on building the spine from places identified in Wikidata and Library of Congress: there was also strong interest expressed in beginning with places identified in a world-historical atlas. In the end, it was determined that using a well-established and respected atlas of world history as the initial source for a list of historical place names was to be the most appropriate path.

In addition to the basic spine, work was to proceed on further specification of place names, their geography, time, and relationships. It was clearly emphasized that specialized

gazetteers should be constructed and linked to the spine: most prominent of these is a gazetteer of administrative units (identifying their legal status, boundaries, and hierarchy). Other specialized gazetteers are to focus on specific regions and topics. In addition, there was strong interest in building a web front end (on the CHIA site) to display the gazetteer, plus interest in specifying context (local to global) and in development of a temporal ontology. There was a recommendation to create an API (application program interface) to link the gazetteer with historical data; with time the group understood that the emphasis on linked open data (connection through URIs) meant that transmission of whole datasets by API might not be necessary.

Post-Workshop Activities, 2014-2015

While the great majority of project funding was devoted to the workshop, participants in the workshop were able to carry on additional activities during the project year, as follows:

Report on Gazetteer Data Resources. Kathy Weimer and Tonia Sutherland conducted a survey of several geographic resources (Wikidata, Wikipedia, Geonames, LOC, and others) to establish their relative strengths. A report, completed in December 2014, showed that the varying resources have substantial overlap and significant cross-referencing but that each has its individual approach and some distinctive data. No one of them stood out unmistakably as the basis for a world-historical gazetteer.¹

Spine Gazetteer: Atlas and registry. Based on the data resources report, work went ahead to rely on a world-historical atlas build a spine gazetteer database and registry. *The Atlas of World History*, edited by Jeremy Black, has a detailed index that project staff scanned in January and February. The result yielded over ten thousand places, each with temporal references and other data. The gazetteer is stored on the stage (internal) site of the CHIA website at stage.chia.pitt.edu, and its interface is in development. We distinguish the “simple spine” (places only) from the “documented spine.” In the latter we are adding structured disambiguation relationships, to add documentation when possible by name, location, and time, drawn from the atlas and other resources. For this purpose, Leif Isaksen, Rainer Simon, and Tom Elliott will make available to this project the principal tools with which they have worked in the Pelagios project. This development has gone beyond the initial goals of the DH Start-up project and was funded by support from the University of Pittsburgh. The major obstacle that must be overcome in order to achieve further gazetteer development centers on the rights to utilize the Black atlas as the primary or sole source for the spine Gazetteer. Until we receive permission from the atlas publisher, or until we determine that the current copyright laws do not apply in this instance, the spine cannot be made publically available.

Encompassing vague and specific documentation of places. The gazetteer, to address the range of world-historical places over the past several centuries, must be able to hold and display multiple levels of documentation, so as to link fully-documented contemporary places with earlier places for which latitude and longitude are not certain for their location or their boundaries. African places were selected as the best region for addressing this issue, in that they were studied under steadily shifting techniques in the nineteenth and early twentieth centuries. Matt Drwenski began work on this issue, under the guidance of

¹ Results of the report were included in a published research report: Tonia Sutherland, “Complexities in Creating and Defining Topical Metadata: Practices for a World-Historical Data Resource,” *Journal of World-Historical Information* 2-3 (2015), doi 10.5195/jwhi.2015.22.

Humphrey Southall, through coding of places in Edward Hertslet, *The Map of Africa by Treaty*, 3rd ed., 3 vols. (London: Cass, 1967). This project, while not completed during the project year, showed itself to be feasible and valuable.

The World-Historical Gazetteer and the larger CHIA Project. The World-Historical Gazetteer was conceived within the larger Collaborative for Historical Information and Analysis, intended to create a comprehensive world-historical data resource, applying it initially to the study of global inequality. The DH Start-up project on the gazetteer brought several advances to the larger CHIA project and in turn benefited from developments elsewhere in the CHIA project.

The complexity of geographic specification, as we discovered it through work on the gazetteer, led us to see that overall documentation needed to be conducted at two levels. Level 1 documentation consists of 20 fields documenting sources and variables, but it has space only for listing the maximum geographic extent for each dataset. For the full implementation of the gazetteer, noting the specific places included in each dataset, a more elaborate Level 2 documentation is required. More broadly, the developing design for linkage among the spine gazetteer and other, specialized gazetteers clarified the design of metadata organization for the CHIA project as a whole, including plans for documenting time, topic, and scale.

In turn, developments elsewhere in the CHIA project strengthened work on the World-Historical Gazetteer. The Dataverse Archive on which CHIA's public data are held, based at Harvard University, was upgraded to Dataverse 4.2. The CHIA website (chia.pitt.edu) was expanded to incorporate all aspects of the project, including the gazetteer. Further, the CHIA data submission system was upgraded to a single path for data ingest. Finally, the infrastructure of Col*Fusion was upgraded to version 2.0. This linkage infrastructure provides the platform on which analysis is to take place—it facilitates documentation, comparison, key-word search and matching. Altogether, these changes in the CHIA environment strengthened the framework for our gazetteer. Further versions of Col*Fusion will link and aggregate datasets to connect integrated data from Col*Fusion to the linked-data cloud.²

Facing the issues and making the decisions in the gazetteer project ultimately removed blockages in the overall CHIA project. It led to a realization that inconsistent dataset documentation was a major obstacle to CHIA's work, to creation of the specifics of Level 1 documentation and to the general design of Level 2 documentation. The question of how to aggregate files within CHIA was debated in discussions at the fringe of the gazetteer conference. This discussion, peripheral to the gazetteer but central to the larger CHIA project, opposed proponents of building one huge dataset of all incorporated files to those who favored aggregating through links among distributed datasets. Provisionally, both paths are to be followed until we can compare results from each. One more point, central to the future of the gazetteer, became clear: the next step of gazetteer work involves modeling and implementing links of the spine gazetteer to locally or topically specific gazetteers, to establish the specifics of complementarity and differentiation in this distributed ontological resource. This objective will be the focus of our future funding proposals, listed below.

² Evgeny Karataev and Vladimir Zadorozhny, "Col*Fusion: A global-scale information integration infrastructure based on crowdsourcing," *Journal of World-Historical Information* 2-3 (2015), doi 10.5195/jwhi.2015.25.

Accomplishments and Evaluation

Accomplishments. The first major accomplishment of the project was the assembly of an interdisciplinary team of major researchers, knowledgeable about the needs, criteria, design, and construction of a world-historical gazetteer. The second accomplishment was an intensive two-day workshop that achieved basic consensus on design for a gazetteer and a basic plan for its construction. Further, the team listed and reviewed the principal existing resources for the gazetteer. On this basis, the project then created the “spine” gazetteer of over 10,000 places of world-historical relevance out of a major historical atlas and arranged for hosting this initial stage of the gazetteer on the stage website of CHIA website, where it is nearly ready for release. We also encountered and overcame differences in approaches that appeared among project participants after the initial workshop.

Evaluation. The workshop was evaluated through circulation of a summary of workshop activities and decisions, followed by active online commentary by participants. The co-PIs and the Pittsburgh staff then identified new questions and choices on gazetteer design that emerged after the workshop. One of these was clarification of distinction between the general, “spine” gazetteer and the gazetteer of administrative units. In addition, development of the gazetteer facilitated discussion of important issues in the larger CHIA world-historical data resource, such as the question of the format of the aggregated datasets that are to be created. Construction of the spine gazetteer based on a world-historical atlas and its installation on the CHIA website resolved many of the problems at this stage and clarified the steps that need to be taken for further development.

Future Work: Articulation of Spine Gazetteer with Other Resources

While this start-up project has achieved essential steps in the creation of a World-Historical Gazetteer for the period since 1500 CE, there is much work to be done in order to link this basic spine with other gazetteer resources and to develop the fully documented version. For further development and implementation, we have identified five major types of work, each of which requires articulation of the spine gazetteer with other resources. We find that the development of the World-Historical Gazetteer in company with the creation of a world-historical data resource now poses a higher-order challenge on the linkage of data.

The first goal in future development is that of progressive integration of the gazetteer into the larger CHIA system of historical information and analysis. Through advancing systems of ingest, the CHIA infrastructure will contribute to the analysis and export of spatially documented files that will be added to the already existing database of place names. The gazetteer will serve as a resource or code book to be applied to each dataset at the point of ingest to fully describe the spatial data contained within.

The second area of work is to add additional functionality to the gazetteer. Each place name will need to be set in spatial context from a local to a global level. Additionally, a graphic with open annotation allowing the traversing of space and will include a function for nongazetteer referencing of attestations of places. More generally, the task is to link the spatial gazetteer to parallel ontologies of time, topic, and scale that are to be developed for historical data. In order for this goal to be realized, it may be necessary to develop a specific post within the existing CHIA staff, to address the linkage of ontologies across the multiple dimensions of historical data.

Thirdly, substantial effort will go into developing system of interaction and mutual

information of the spine gazetteer with more specific gazetteers. Among these, the spine gazetteer for places generally must be distinguished from but linked to a gazetteer for administrative units. The key figure in emphasizing this distinction is Humphrey Southall, director of the PastPlace project, who has laid out the strategy for a gazetteer of administrative units, names, their relationships, the polygons of their edges, and the statistics they include. In addition, the spine gazetteer must be linked to regional and topical gazetteers to establish a distributed system of global spatial information. The principal historical research projects that we see as ready for building specialized gazetteers are (1) the studies of CLIO-INFRA (at IISH in Amsterdam) on wages and prices in Europe and elsewhere; (2) the data on life transitions in slave societies of the Americas (at ESSSS at Vanderbilt); and (3) the Asia-wide data on the 1918 influenza pandemic (Asian Studies at Michigan State). The linking of the spine gazetteer to administrative units and regional/topical gazetteers will be a primary focus of future funding proposals.

The fourth major area of work is in expanding and deepening the human collaboration within each participating work group and among the groups. This requires regular meetings that bring together participants with different specializations. One example of such collaboration is the handling of distributed archives. CHIA seeks to hold and display many datasets in order to ensure that their spatial documentation is consistent. It is recognized, however, that many individual scholars and institutions will wish to maintain ownership and copies of their data. The point is to develop a distributed archive making it possible to own and share data at the same time. Maintenance of consistent documentation will help ensure that users of datasets give consistent citation of the owners and creators.

Finally, as the simple spine and the documented spine develop, along with their links to specialized gazetteers, there will be a need to test and advance the reliability of this expanding World-Historical Gazetteer in providing consistent spatial documentation. This process will be ongoing throughout the development of the gazetteer and after its implementation. In addition to utilizing existing resources such as the ingested entries in the English-language Wikipedia with geographical coordinates, we will develop large-scale collections of data from historical datasets contained in the CHIA collection and from names harvested from print historical documents. Further extension of the gazetteer will ensure that it addresses texts in multiple languages. These resources will be used as testbeds upon which we will measure the reliability and coverage of the simple and documented spine.

Future proposals

The CHIA project has identified two major sources for funding to expand and implement the work begun on the World-Historical Gazetteer project. Along with preparing an application for a Digital Humanities Implementation grant from the NEH to continue the work initiated by the World-Historical Gazetteer Workshop in 2014, we are currently developing a proposal for the National Science Foundation's Resource Implementation for Data Intensive Research (RIDIR) grant competition. While the NEH proposal will focus strictly on the further development of the World-Historical Gazetteer as described in this paper, the proposal to NSF-RIDIR will look to fund more broadly the development of a next-generation data resource of which the gazetteer will be a major component. The

developed databases and infrastructure should have significant impact across multiple fields in the social sciences and lead to new types of data-intensive and cross-disciplinary research.

Appendix 1: Workshop Agenda and Participant List

Thursday, September 4th

- 8:30 am Greeting and Continental Breakfast
- 9:00 am Opening remarks: Patrick Manning
- 9:25 am Opening remarks: Ruth Mostern
- Introduction of upcoming book *Placing Names: Enriching and Integrating Gazetteers* edited by Mostern, Southall, and Berman
- 9:45 am The DH 2014 conference report from Karl Grossner and Kathy Weimer
- Developments from the Digital Humanities Conference in Lausanne
- 10:30 am The Pelagios Project: Leif Isaksen
- Pelagios: Enable Linked Ancient Geodata In Open Systems
- 11:30 am The technical side of Pelagios: Rainer Simon
- Cross-gazetteer search framework and gazetteer interlinking RDF profiles
- 12:15 pm Past Place - a prototype world-historical gazetteer: Humphrey Southall
- Hands-on demo of Past Place in comparison to other gazetteers
- 1:00 pm Lunch
- 2:00 pm Small Group Sessions
- Proposed topic for discussion: “Developing networks of collaboration for creating a Gazetteer system with a single set of shared IDs post 1500 C.E.”
- 4:00 pm Continuation of Small Group Session
- Refined discussion focusing on inventory, linking data, standards, and ingest
- 5:00 – 5:30 pm Small Group Reports
- 6:30 pm Dinner

Friday, September 5th

- 9:00 am Morning Remarks: Patrick Manning
- 9:15 am Video montage: “How do you build a World-Historical Gazetteer?”
- Thought of project participants recorded during the day on Thursday
- 9:30 am View from the Social Scientists
- Short presentations from Chandra, Gerring, Mostern, Krishna, and Fall on how a World-Historical Gazetteer would be used in social science research
- 11:15 am Small Group Reconvene for Summary Discussion
- Points of agreement, identification of problems, moving forward
- 12:00 pm Bringing It Together
- Moderated session with goal of creating a conclusive plan of action
- 1:00 pm Lunch
- 2:00 pm Plan Implementation
- Moderated session on how best to implement the plan of action
- 3:45 pm Moving Forward
- Linking plan implementation to specific future focused actions
- 5:00pm Closing Remarks from Patrick Manning and Ruth Mostern

Appendix 2. Project Participants

Micah Altman Director of Research and Head/Scientist in the Program on Information Science for the MIT Libraries at the Massachusetts Institute of Technology

Marieka Arksey Ph.D. student in the World Cultures Interdisciplinary program at University of California, Merced

Merrick Lex Berman Manager of Chinese and Japanese geographic information projects such as

CHGIS, the Skinner Archive, and JapanMap at Harvard University

Aaron Brenner Coordinator of Digital Scholarship in the University Library System at the University of Pittsburgh

Siddharth Chandra Director of the Asian Studies Center at Michigan State University and an expert in international economics and international affairs

Matt Drwenski Graduate Student in World History at University of Pittsburgh focusing on Modern World History, Globalization, Economic History, and World-systems theory

Tom Elliott Associate Director for Digital Programs and Senior Research Scholar at New York University focusing on the practice of digital humanities in ancient studies

Mamadou Fall Professor of History at the Université Cheikh Anta Diop, Dakar, author of recent major work on historical geography of Senegambia, with emphasis on “terroir” as a socially defined space

John Gerring Professor of Political Science at Boston University with backgrounds in methodology and comparative politics and director of the CLIO World Tables

Karl Grossner Geographer and digital humanities research developer at Stanford currently working to incorporate spatial and temporal computing methods into long-term humanities projects

Leif Isaksen Lecturer in Archaeology and Digital Humanities and Deputy Director of the Web Science Doctoral Training Centre at the University of Southampton

Evgeny Karataev Doctoral Student at University of Pittsburgh in the School of Information Sciences and developer of Col*Fusion data ingest and linkage system with interests in complex adaptive information systems

Matthew Knutzen Geospatial librarian and Assistant Chief of the Map Division at New York Public Library

Shekhar Krishnan Social scientist and consultant with a PhD in History and Anthropology of Science and Technology from Massachusetts Institute of Technology

Patrick Manning Andrew W. Mellon Professor of World History, Director of the World History Center at the University of Pittsburgh, Director of the CHIA Project, and 2016 President of the American Historical Association

Ruth Mostern Associate Professor in the School of Social Sciences, Humanities and Arts with a focus on historical geography and former head of Collection Development at the Electronic Cultural Atlas Initiative

Rainer Simon Scientist at the Austrian Institute of Technology with a background in semantic technologies and Linked Data in the Digital Humanities and Digital Libraries fields and Technical Director of the Pelagios Project

Raj Singh Former Director of Interoperability Programs for the Open Geospatial Consortium

Humphrey Southall Professor in Historical Geography at Portsmouth University and Director of the Great Britain Historical GIS Project

Tonia Sutherland Postdoctoral Associate at the World History Center at the University of Pittsburgh with a PhD from the School of Information Sciences at Pitt

Kathy Weimer Professor and Curator of Maps at Texas A&M University and Coordinator for the Map and GIS Library

Vladimir Zadorozhny Associate Professor of Graduate Information Science and Technology Program at the School of Information Sciences at Pitt with interest in networked information systems, integration, and data fusion